

Space-Optimal Semi-Streaming for $(2 + \varepsilon)$ -Approximate Matching

Mohsen Ghaffari
ETH Zurich
ghaffari@inf.ethz.ch

Abstract

In a recent breakthrough, Paz and Schwartzman [SODA'17] presented a single-pass $(2 + \varepsilon)$ -approximation algorithm for the maximum weight matching problem in the semi-streaming model. Their algorithm uses $O(n \log^2 n)$ bits of space, for any constant $\varepsilon > 0$.

In this note, we present a different analysis, for essentially the same algorithm, that improves the space complexity to the optimal bound of $O(n \log n)$ bits, while also providing a more intuitive explanation of the process. This space complexity is optimal because just keeping the matching needs $\Omega(n \log n)$ bits.

1 Introduction and Related Work

The maximum weight matching (MWM) problem is a classical optimization problem, with diverse applications, which has been studied extensively since the 1965 work of Edmonds [Edm65]. Naturally, this problem has received significant attention also in the *semi-streaming* model. This is a modern model of computation, introduced by Feigenbaum et al. [FKM⁺05], which is motivated by the need for processing massive graphs whose edge set cannot be stored in memory.

There has been a sequence of successively improved approximation algorithms for MWM in the semi-streaming model. Feigenbaum et al. gave a 6 approximation [FKM⁺05], McGregor gave a 5.828 approximation¹ [McG05], Epstein et al. gave a $4.911 + \varepsilon$ approximation [ELMS11], Crouch and Stubbs gave a $4 + \varepsilon$ approximation [CS14], and Grigorescu et al. improved the bound to 3.5 [GMZ16]. However, these approximations remained far away from the more natural and more familiar 2 approximation, which the sequential greedy method provides.

In a recent breakthrough, Paz and Schwartzman [PS17] presented an extremely simple algorithm that achieves an approximation of $2 + \varepsilon$. Their algorithm uses $O(n \log^2 n)$ bits of space. More concretely, they maintain $O(n \log n)$ edges, while working through the stream. Then at the end, they compute the matching based on these maintained edges.

Our contribution: We present an alternative analysis for (a simple variant of) their algorithm, which has two advantages: (1) it implies that keeping merely $O(n)$ edges suffices, and thus improves the space complexity to $O(n \log n)$ bits, which is optimal, (2) it provides a more intuitive explanation for the strength of this simple method.

Roadmap: In Section 2, as a warm up, we review (a simple version of) the algorithm of Paz and Schwartzman. Moreover, we present an alternative and more intuitive style of analysis for it, which will be used also in our improvement. In Section 3, we present the improved algorithm/analysis that achieves a space complexity of $O(n \log n)$.

¹He also presented a $2 + \varepsilon$ approximation, but using $O(1/\varepsilon^3)$ passes on the input.

2 Reviewing the Algorithm of Paz and Schwartzman

The setting: Let $G = (V, E, w)$ be a simple graph with non-negative edges weights $w : E \rightarrow \mathbb{R}_{\geq 0}$. Let $n = |V|$, $m = |E|$. Without loss of generality, suppose that the edge weights are normalized so that the minimum edge weight is 1 and let W denote the maximum edge weight.

In the semi-streaming model, the input graph G is provided as a stream of edges. In each iteration, the algorithm receives an edge from the stream and processes it. The algorithm has a memory much smaller than m and thus it cannot store the whole graph. The amount of the memory that the algorithm uses is called its *space complexity* and we wish to have it as small as possible. The objective of the algorithm in the *maximum weight matching* (MWM) problem is that, at the end of the stream, the algorithm outputs a matching, whose weight is close to the weight of the maximum weight matching.

2.1 The Basic Algorithm

The starting point in the approach of Paz and Schwartzman [PS17] is the following basic yet elegant algorithm. For the sake of explanation, consider a sequential model of computing. We later discuss the adaptation to the streaming model.

Basic Algorithm: Repeatedly select an edge e with positive weight; reduce its weight from itself and its neighboring edges; push e into a stack and continue to the next edge, so long as edges with positive weight remain. At the end; unwind the stack and add the edges greedily to the matching.

In the following, we argue that this simple algorithm computes a 2-approximation of the maximum weight matching. We present an analysis that is morally equivalent with the proof given in [PS17] based on the *local ratio theorem*, but is put in a more intuitive language. Moreover, the same analysis style will be much more convenient for the explanation of the improvement.

Lemma 2.1. *The matching M returned by the basic algorithm is a 2-approximation of the maximum weight matching.*

Proof. Consider an arbitrary matching M' . We prove that there exists a way of moving the weight of the M' -edges onto M -edges such that each M -edge e receives at most $2w(e)$ weight. That is, we start with $w(e')$ dollars on each M' -edge e' , and we will move around this money such that at the end the money is only on M -edges and each M -edge e has at most $2w(e)$ dollars. Since the total money is always kept the same, this directly implies that $\sum_{e' \in M'} w(e') \leq 2 \sum_{e \in M} w(e)$. In a sense, we are *blaming* the fact that we did not take M' -edges on M -edges, in a way that each M -edge receives a blame at most twice its weight. Hence, M is a 2-approximation.

Consider the time that we remove an edge $e_1 = (v, v')$ and put e_1 in the stack. We deduct $w(e_1)$ from the weight of all edges incident on v or v' . We are then left with a new graph G_r with updated weights w_r . On the rewind of the stack, when processing e_1 , there are two possibilities:

- (A) the matching \mathcal{M}_r computed on G_r has an edge e'' incident on v or v' , or
- (B) the matching \mathcal{M}_r has no edge incident on v or v' , thus we will add e_1 to the matching of G .

Suppose by induction that on G_r , there is a way of blaming the weights w_r of M' -edges on the edges of \mathcal{M}_r so that each \mathcal{M}_r -edge e receives at most $2w_r(e)$ blame. M' contains at most two edges e_2 and e_3 incident on v or v' . Notice that one of these may be e_1 , in which case in fact there is only one of them. By the inductive assumption, for each $e_i \in \{e_2, e_3\}$, we have already found room

for placing the $w_r(e_i) = w(e_i) - w(e_1)$ part of the blame on \mathcal{M}_r edges. Now we need to find room for at most $(w(e_2) - w_r(e_2)) + (w(e_3) - w_r(e_3)) = 2w(e_1)$ more blame. In case (A), edge $e'' \in \mathcal{M}_r$ has at most $2w_r(e'') = 2(w(e'') - w(e_1))$ blame on it at the moment. But in G , edge e'' has room for $2w(e'')$ blame. Hence, in this case, we have room for that $2w(e_1)$ extra blame, to be placed on e'' . In case (B), where we add e_1 to the matching, e_1 is a fresh addition to the matching with zero blame on it so far. Thus, we again have room for placing the $2w(e_1)$ extra blame, this time on e_1 , while ensuring that each edge has at most twice its weight as blame. \square

Implementing the Basic Algorithm in the Semi-Streaming Model: To implement the above algorithm while working through the stream, we just need to remember a parameter $\phi(v)$ for each node v . This parameter is the total sum of the weight already reduced from the edges incident on vertex v , due to edges incident on v that were processed and put in the stack before.

However, the space complexity of this basic algorithm can be quite high. In particular, we may end up pushing even $\Theta(n^2)$ edges into the stack. This brings us to a clever idea of Paz and Schwartzman [PS17], as we discuss next.

2.2 The Algorithm with Exponentially Increasing Weights

We now briefly overview an idea of Paz and Schwartzman [PS17] which cuts the space complexity to the equivalent of keeping $O(n \log W/\varepsilon)$ edges, where W is the normalized maximum edge weight, while still providing a $(2 + \varepsilon)$ approximation. The idea is to ensure that the edges incident on each node v that get pushed into the stack have exponentially increasing weights, by factors of $(1 + \varepsilon)$. Thus, per node at most $O(\log W/\varepsilon)$ edges are added to the stack. This ensures that the overall number of edges in the stack is at most $O(n \log W/\varepsilon)$.

To attain this exponential growth, the method is as follows: When doing a step of reducing the weight of an edge e from each neighboring edge e' , we will decide between deducting either $w(e)$ or $(1 + \varepsilon)w(e)$ from the weight of $w(e')$. In general, this can be any arbitrary decision. In the streaming model, this decision is done when we first see $e' = \{u, u'\}$ in the stream, as follows.

- If $w(e') \leq (1 + \varepsilon)(\phi(u) + \phi(u'))$ — i.e., if e' has less than $(1 + \varepsilon)$ times of the total weight of the stacked up edges incident on u or u' — then we deduct $(1 + \varepsilon)w(e)$ from $w(e')$ for each stacked up edge e incident on u or u' . Hence, effectively we reduce the weight of $w(e')$ by $(1 + \varepsilon)(\phi(u) + \phi(u'))$. This implies that we get left with an edge e' of non-positive weight, which can be ignored without putting in the stack.
- Otherwise, if $w(e') \geq (1 + \varepsilon)(\phi(u) + \phi(u'))$, we deduct only $w(e)$ from $w(e')$, for each edge e incident on u or u' that is already in the stack. Thus, in total, we deduct only $(\phi(u) + \phi(u'))$ weight from $w(e')$ for the previously stacked edges. Thus, now we have an edge e' whose leftover weight is $w'(e') \geq (1 + \varepsilon)(\phi(u) + \phi(u')) - (\phi(u) + \phi(u')) = \varepsilon(\phi(u) + \phi(u'))$. Then, we add e' to the stack, and thus $\phi(u)$ and $\phi(u')$ increase by $w'(e')$. Therefore, each of $\phi(u)$ and $\phi(u')$ increases by at least a $(1 + \varepsilon)$ factor.

The concrete algorithm that formalizes the above scheme is presented in Line 14.

Observation 2.2. *When an edge $e = \{u, u'\}$ gets added to the stack, the value of $\phi(u)$ increases by a $1 + \varepsilon$ factor.*

The above observation also implies that the edges incident on each node that are added to the stack have an exponential growth in weight.

Lemma 2.3. *The matching M returned by Line 14 is a $2(1 + \varepsilon)$ approximation of the maximum weight matching.*

Algorithm 1 The Algorithm of Paz-Schwartzman [PS17] With Exponentially Increasing Weights

```

1:  $S \leftarrow \text{emptystack}$ 
2:  $\phi = \text{zeros}(1, n)$ 
3: for  $e = (u, u') \in E$  do
4:   if  $w[e] \leq (1 + \varepsilon)(\phi(u) + \phi(u'))$  then
5:     continue ▷ skip to the next edge
6:    $w'[e] \leftarrow w[e] - (\phi(u) + \phi(u'))$ 
7:    $\phi(u) \leftarrow \phi(u) + w'[e]$ 
8:    $\phi(u') \leftarrow \phi(u') + w'[e]$ 
9:    $S.\text{push}(e)$ 

10:  $M \leftarrow \emptyset$ 
11: while  $S \neq \emptyset$  do
12:    $e \leftarrow S.\text{pop}()$ 
13:   if  $M \cap N(e) = \emptyset$  then  $M \leftarrow M \cup \{e\}$ 
14: return  $M$ 

```

Proof. The proof is similar to that of Lemma 2.1, so we only discuss the necessary changes. Following our argument of blaming the weight of an arbitrary matching M' on the edges of the computed matching M , the blaming will now be such that each M -edge e receives at most $2(1 + \varepsilon)$ times its weight as blame. Hence, the computed matching is a $2(1 + \varepsilon)$ approximation.

Consider the time that we remove an edge $e_1 = (v, v')$ and put e_1 in the stack. We now may deduct either $w(e_1)$ or $(1 + \varepsilon)w_e$ from each of the edges incident on v or v' . We then get left with a new graph G_r with updated weights w_r . On the rewind of the stack, we have two possibilities as before: either (A) the matching \mathcal{M}_r computed on G_r has an edge e'' incident on v or v' , or (B) we will add e_1 to the matching of G .

Suppose by induction that on G_r , there is a way of blaming the weights w_r of M' -edges on the edges of \mathcal{M}_r , in a manner that each \mathcal{M}_r -edge e receives at most $2(1 + \varepsilon)w_r(e)$ blame. M' contains at most two edges e_2 and e_3 incident on v or v' . By the inductive assumption, for each $e_i \in \{e_2, e_3\}$, we have already found room for placing the $w_r(e_i) \geq w(e_i) - (1 + \varepsilon)w(e_1)$ part of the blame on \mathcal{M}_r edges. Now we need to find room for at most $(w(e_2) - w_r(e_2)) + (w(e_3) - w_r(e_3)) \leq 2(1 + \varepsilon)w(e_1)$ more blame. In case (A), edge $e'' \in \mathcal{M}_r$ has at most $2(1 + \varepsilon)w_r(e'') \leq 2(1 + \varepsilon)(w(e'') - w(e_1))$ blame on it at the moment. But in G , edge e'' has room for $2(1 + \varepsilon)w(e'')$ blame. Hence, in this case, we have room for that $2(1 + \varepsilon)w(e_1)$ extra blame, to be placed on e'' . In case (B), where we add e_1 to the matching, e_1 is a fresh addition to the matching with zero blame on it so far. Thus, we again have room for placing the $2(1 + \varepsilon)w(e_1)$ extra blame, this time on e_1 , while ensuring that each edge has at most $2(1 + \varepsilon)$ times its weight as blame. \square

3 Improved Algorithm/Analysis

The algorithm presented in the previous section maintains $O(n \log W)$ edges. To improve the space complexity, we would like to keep only $O(n)$ edges. For that purpose, we will limit the number of edges incident on each vertex v that are in the stack to a constant $\beta = \frac{5 \log 1/\varepsilon}{\varepsilon}$. When there are more edges, we'll take out the earliest one and remove it from the stack. This will be easy to implement using a queue $Q(v)$ for each of vertex v , where we keep the length of the $Q(v)$ capped to β . The pseudo-code is presented in Line 20. We will prove in the following that this cannot hurt the approximation factor more than just increasing ε by a 2 factor.

Remark: Paz and Schwartzman [PS17] used a similar algorithmic idea to keep only $O(n \log n)$

Algorithm 2 The Optimal-Space Algorithm

```

1:  $S \leftarrow \text{empty stack}$ 
2:  $\forall v \in V : Q(v) \leftarrow \text{empty queue}$ 
3:  $\phi = \text{zeros}(1, n)$ 
4: for  $e = (u, u') \in E$  do
5:   if  $w[e] \leq (1 + \varepsilon)(\phi(u) + \phi(u'))$  then
6:      $\text{continue}$   $\triangleright$  skip to the next edge
7:    $w'[e] \leftarrow w[e] - (\phi(u) + \phi(u'))$ 
8:    $\phi(u) \leftarrow \phi(u) + w'[e]$ 
9:    $\phi(u') \leftarrow \phi(u') + w'[e]$ 
10:   $S.\text{push}(e)$ 
11:  for  $v \in \{u, u'\}$  do
12:     $Q(v).\text{enqueue}(e)$ 
13:    if  $|Q(v)| \geq \beta = \frac{5 \log 1/\varepsilon}{\varepsilon}$  then
14:       $e' \leftarrow Q(v).\text{dequeue}()$ 
15:       $\text{remove } e' \text{ from the stack } S$ 

16:  $M \leftarrow \emptyset$ 
17: while  $S \neq \emptyset$  do
18:    $e \leftarrow S.\text{pop}()$ 
19:   if  $M \cap N(e) = \emptyset$  then  $M \leftarrow M \cup \{e\}$ 
20: return  $M$ 

```

edges in total, instead of $O(n \log W)$ edges. To be precise, they keep $\gamma = \Theta(\log n / \varepsilon)$ edges per node. The difference and thus the improvement lies in the analysis. In a rough sense, their argument was that, per step, the process of limiting the queue size to γ creates a loss of $(1 - \exp(-\gamma))$ factor in the approximation. Thus, over all the up to $O(n^2)$ edges in the stream, the loss is $(1 - \exp(-\gamma))^{O(n^2)}$. This is why they had to set $\gamma = \Theta(\log n)$ to make the loss negligible. We will use a different way of accounting for the loss, continuing with our *blaming* argument, which will allow us to curtail the per-node queue size to $\beta = O(1)$, while keeping the loss negligible.

Observation 3.1. *Suppose that an edge $e = \{v, u\}$ in the stack gets removed from the stack because another edge $e' = \{v, u'\}$ was pushed to the stack later and made the size of the queue $Q(v)$ reach $\beta + 1$. Then, we say e' evicted e . The left-over weight $w'(e')$ at the time of inserting e' in the stack is at least a $1/\varepsilon^3$ factor of the left-over weight $w'(e)$ at the time of the insertion of e to the stack.*

Proof. Since e was evicted by e' , there must have been $\beta - 1$ edges incident on v that arrived after e and before e' and were pushed into the stack. Hence, because of Observation 2.2, at the time of the arrival of e' , we have $\phi(v) \geq (1 + \varepsilon)^{\beta-1} w'(e) \geq w'(e)/\varepsilon^4$. But since e' got added to the stack, we know that $w'(e') \geq \varepsilon \phi(v) \geq w'(e)/\varepsilon^3$. \square

Lemma 3.2. *The matching M returned by Line 20 is a $2(1 + 2\varepsilon)$ approximation of the maximum weight matching.*

Proof. Again, we will continue with our style of *blaming*, as in Lemma 2.1 and Lemma 2.3. We now blame the weight of any arbitrary matching M' on the computed matching M , in a manner that each M -edge receives at most $2(1 + 2\varepsilon)$ factor of its weight as blame.

The only difference in the algorithm is that now some edges are thrown out of the stack, because the queue size of one of their endpoints grew larger than β . For the sake of analysis, let us assume that these thrown out edges are still kept in the stack, and we see them when we unwind the stack, but we cannot add them to the matching (as they are forgotten).

Let us consider one edge $e_1 = \{v, u\}$ that was added to the stack, but later thrown out because another edge $e'_1 = \{v, u'\}$ arrived that got added to the stack. That is, e'_1 evicted e_1 . Notice that, as in the argument of Lemma 2.3, all that we need to find is room for placing $2(1 + \varepsilon)w'(e_1)$ blame on some edges of the computed matching M . When processing e_1 on the rewind of the stack, there are two possibilities, similar to before: The case (A) where there is an edge e'' neighboring e_1 that is in the matching would be same as before and we would place this extra blame on e'' . But the case (B) where there is no edge neighboring e_1 in the matching is different, because we now cannot add e_1 to the matching. This is because e_1 is effectively forgotten (from the stack). Hence, we also cannot put the $2(1 + \varepsilon)w'(e_1)$ extra blame on edge e_1 . In this case, we will need to find room for this extra blame elsewhere, as we explain next.

The reason that e_1 got evicted is edge e'_1 that arrived later, got added to the stack, and thus effectively throw e_1 out. We will trace e'_1 and look for room for the blame through e'_1 . Now, it is possible that e'_1 itself also got evicted by some other later arriving edge e'_2 neighboring e'_1 . In fact, this eviction chain may continue for long. Suppose that the maximal eviction chain is e'_1, e'_2, \dots, e'_k where e'_i got evicted because of e'_{i+1} , but e'_k was not evicted from the stack. We will find room for the blame $2(1 + \varepsilon)w'(e_1)$ of e_1 when processing e'_k during the unwinding of the stack. At the time of processing e'_k during the unwinding of the stack, we are looking for $2(1 + \varepsilon)w'(e'_k)$ room for blame for edge e'_k itself. But there is now need for more room for blame, for edge e_1 and other edges similar to e_1 whose eviction chain ends in e'_k .

Notice that there can be many edges whose eviction chain ends in e'_k , even up to $\text{poly}(n)$ many edges. The reason is that each edge may evict two edges, one from each of its endpoints, each of them may have evicted two edges, and so on. Thus the eviction dependencies can be a binary tree of depth $\log_{1+\varepsilon} W = \Theta(n)$. However, as we argue next, the total weight of this tree cannot be large.

Edge e'_k has at most 2 edges that it directly evicted, at most 4 edges with eviction chain of length 2, and more generally at most 2^i edges with eviction chain length of 2^i . On the other hand, thanks to Observation 3.1, the evicted weights are rapidly dropping along these chains. In particular, the weight of each of those edges directly evicted by e'_k is at most $w'(e'_k) \cdot \varepsilon^3$, the weight of each of those with eviction chain of length 2 is at most $w'(e'_k) \cdot \varepsilon^6$, and more generally, the weight of each of those with eviction chain of length i is at most $w'(e'_k) \cdot (\varepsilon)^{3i}$. Hence, the total weight of the eviction tree rooted in e'_k is at most $\sum_{i=1} w'(e'_k) \cdot (2\varepsilon^3)^i \leq w'(e'_k)\varepsilon^2$.

The above implies that, at the time of processing edge e'_k that was not evicted when unwinding the stack, aside from the $2(1 + \varepsilon)w'(e'_k)$ blame room that we need to find for e'_k , we need also $2(1 + \varepsilon)\varepsilon^2w'(e'_k)$ more room for blame, to pay for all those other edges that got evicted because of eviction chains ending in e'_k . Hence, in total, we now need room for at most $2(1 + \varepsilon)w'(e'_k) + 2(1 + \varepsilon)\varepsilon^2w'(e'_k) \leq 2(1 + 2\varepsilon)w'(e'_k)$ blame, when processing e'_k on the unwinding of the stack.

Notice that, since we are aiming for establishing a $2(1 + 2\varepsilon)$ -approximation, we have room for a blame on each edge that is $2(1 + 2\varepsilon)$ times its weight. If we are in case (A) when processing e'_k on the unwind of the stack, meaning that an edge e'' incident on e'_k is already in the matching, then returning the weight of e'' to its weight before the reduction of e'_k opens up space for at least $2(1 + 2\varepsilon)(w(e'')) - 2(1 + 2\varepsilon)(w(e'') - w'(e'_k)) \geq 2(1 + 2\varepsilon)w''(e'_k)$ more blame. Similarly, in case (B), where e'_k itself goes into the matching, there is $2(1 + 2\varepsilon)w''(e'_k)$ room for blame on e'' . Thus, in either case, we have room for blame for e'_k as well as all those edges that were put in the stack but later evicted, with an eviction chain ending in e'_k . \square

Acknowledgment: I am grateful to Gregory Schwartzman for sharing a write-up of [PS17], and to Ami Paz and Gregory Schwartzman for feedback on a draft of this note.

References

- [CS14] Michael Crouch and Daniel M Stubbs. Improved streaming algorithms for weighted matching, via unweighted matching. *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, page 96, 2014.
- [Edm65] Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of mathematics*, 17(3):449–467, 1965.
- [ELMS11] Leah Epstein, Asaf Levin, Julián Mestre, and Danny Segev. Improved approximation guarantees for weighted matching in the semi-streaming model. *SIAM Journal on Discrete Mathematics*, 25(3):1251–1265, 2011.
- [FKM⁺05] Joan Feigenbaum, Sampath Kannan, Andrew McGregor, Siddharth Suri, and Jian Zhang. On graph problems in a semi-streaming model. *Theoretical Computer Science*, 348(2):207–216, 2005.
- [GMZ16] Elena Grigorescu, Morteza Monemizadeh, and Samson Zhou. Streaming weighted matchings: Optimal meets greedy. *arXiv preprint arXiv:1608.01487*, 2016.
- [McG05] Andrew McGregor. Finding graph matchings in data streams. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 170–181. Springer, 2005.
- [PS17] Ami Paz and Gregory Schwartzman. A $(2 + \varepsilon)$ -approximation for maximum weight matching in the semi-streaming model. In *Pro. of ACM-SIAM Symp. on Disc. Alg. (SODA)*, 2017.